Benzo[c]phenanthrene in SRM 1975, Diesel Extract

Mark Vangel

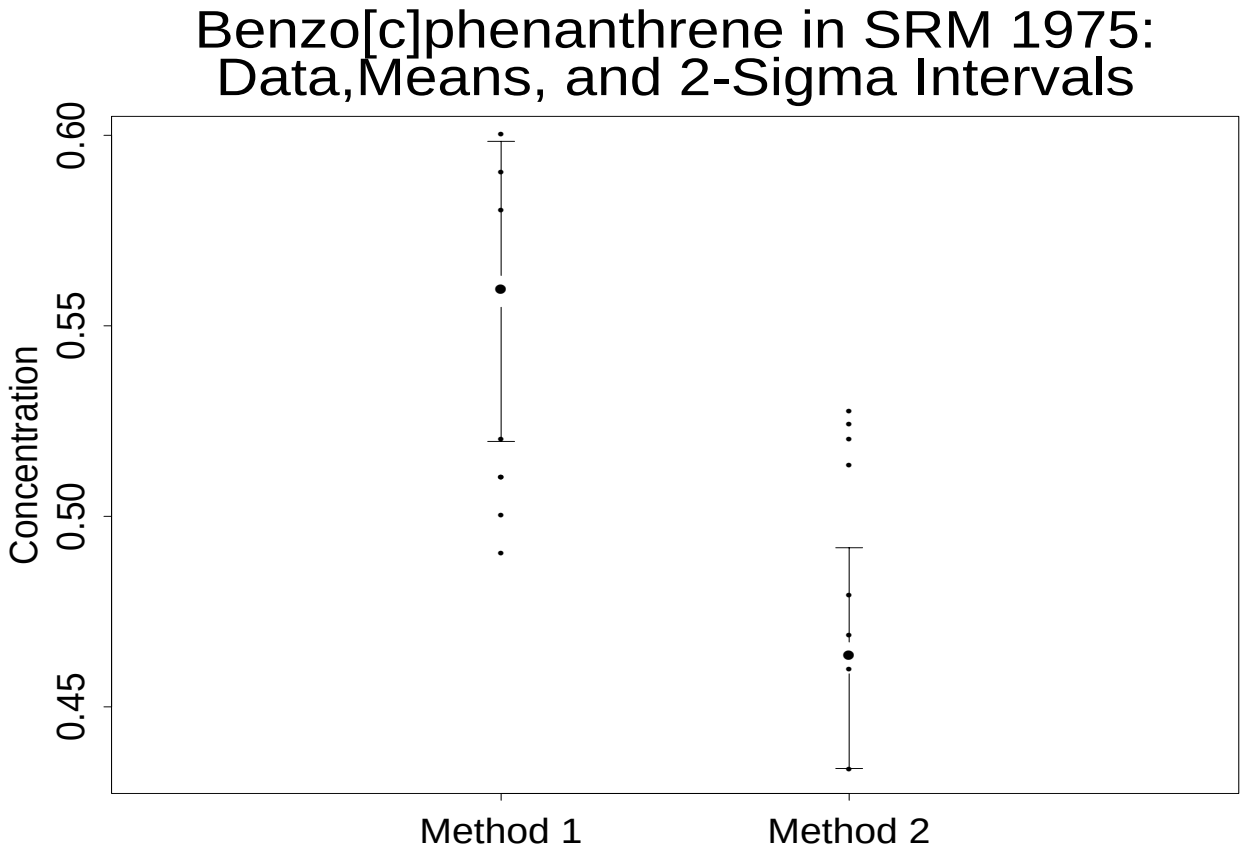
(with additions by J.Yen and K.Kniskern)

The Data

Method	Bottle	Day	Conc.
1	1	1	0.58
1	2	1	0.51
1	3	1	0.59
1	4	1	0.52
1	5	1	0.51
1	6	1	0.49
1	7	1	0.61
1	8	1	0.50
1	9	1	0.60
1	10	1	0.68
2	11	3	0.47
2	12	4	0.43
2	13	2	0.46
2	14	3	0.39
2	15	4	0.40
2	16	2	0.41
2	11	5	0.51
2	12	7	0.52
2	13	5	0.52
2	14	6	0.42
2	15	7	0.48
2	16	5	0.52

This table lists the concentrations of Benzo[c]phenanthrene in SRM 1975, Diesel Extract. How would we use these data to obtain a consensus value and related uncertainties?

Data Plot



This schematic describes how the concentration levels found by Method 1 tend to higher than those found by Method 2. Thus, we cannot treat all the observations as coming from identical distributions.

Outline

We will use the following models to try describe the variation between the results for the different methods:

- Model 1: Individual t-Densities
- Model 2: The 'Behrens-Fisher Problem'
- Model 3: Two-Level Hierarchical Model for Means; One-Level for Variances

On Bayesian Prior and Posterior Distributions

If x_{ij} is the jth observation from the ith sample, then x_{ij} is often modeled as being generated from a distribution $p(x_{ij}|\theta_1, \theta_2, \ldots)$, which is dependent on a set of parameters $\theta_1, \theta_2, \ldots$.

One of the most common used models is the familiar bell-shaped distribution known as the normal distribution:

$$p(x|\mu, \sigma^2) = N(x|\mu, \sigma^2),$$

which stands for a Normal distribution with mean μ and variance σ^2.

Previous knowledge about the parameters is incorporated in prior distributions of that parameter. For example, it may be known that μ has itself the distribution $p(\mu) = N(a, b^2)$ for some $a,\ b$.

For the cases where there is little or no worthwhile prior information, "non-informative" or "objective" priors may be used. A common non-informative prior for the mean is the flat prior $p(\mu) \propto 1$. A common non-informative prior for the standard deviation is $p(\sigma) \propto 1/\sigma$. Note that prior distributions do not have to be proper probability distributions (i.e. they do not have to integrate out to 1).

In a simple case, given the prior distribution of the parameter $\pi(\theta)$ and the data $\{x_i\}$, we can compute the posterior distribution

$$p(\theta|\{x_i\}) = \int p(\{x_i\}|\theta) \cdot \pi(\theta) d\theta.$$

The posterior distribution $p(\theta|\{x_i\})$ encapsulates our best knowledge about θ given both the data and our prior knowledge.

For simpler cases, the posterior distributions can be calculated analytically. For more complicated cases, the posteriors can be generated via Markov Chain Monte Carlo techniques using the freely available BUGS (Bayesian analysis Using Gibbs Sampling) software. A sample BUGS program is on page 18.

BUGS does not work with improper distributions, so a flat prior may be approximated by a normal distribution with a very large variance, and the prior $p(\sigma) \propto 1/\sigma$ can be approximated by a very diffuse Gamma distribution. Note that in BUGS, the Precision $\tau = 1/\sigma^2$ rather than the variance is the input parameter to the Normal (and other) distributions.

Model 1:
Separate Means, Separate Variances

$$\begin{aligned} p(x_{ij}|\theta_i, \sigma_i^2) &= \mathsf{N}(\theta_i, \sigma_i^2) \\ p(\theta_i) &\propto 1 \\ p(\sigma_i) &\propto 1/\sigma_i \end{aligned}$$

Here we model the measurements from the two methods as coming from two different normal distributions. In addition, we have modeled the parameters of the normal distributions with traditional "non-informative" prior distributions (note that prior distributions do not need to be proper probability distributions).

Model 1 Posteriors: t-Densities

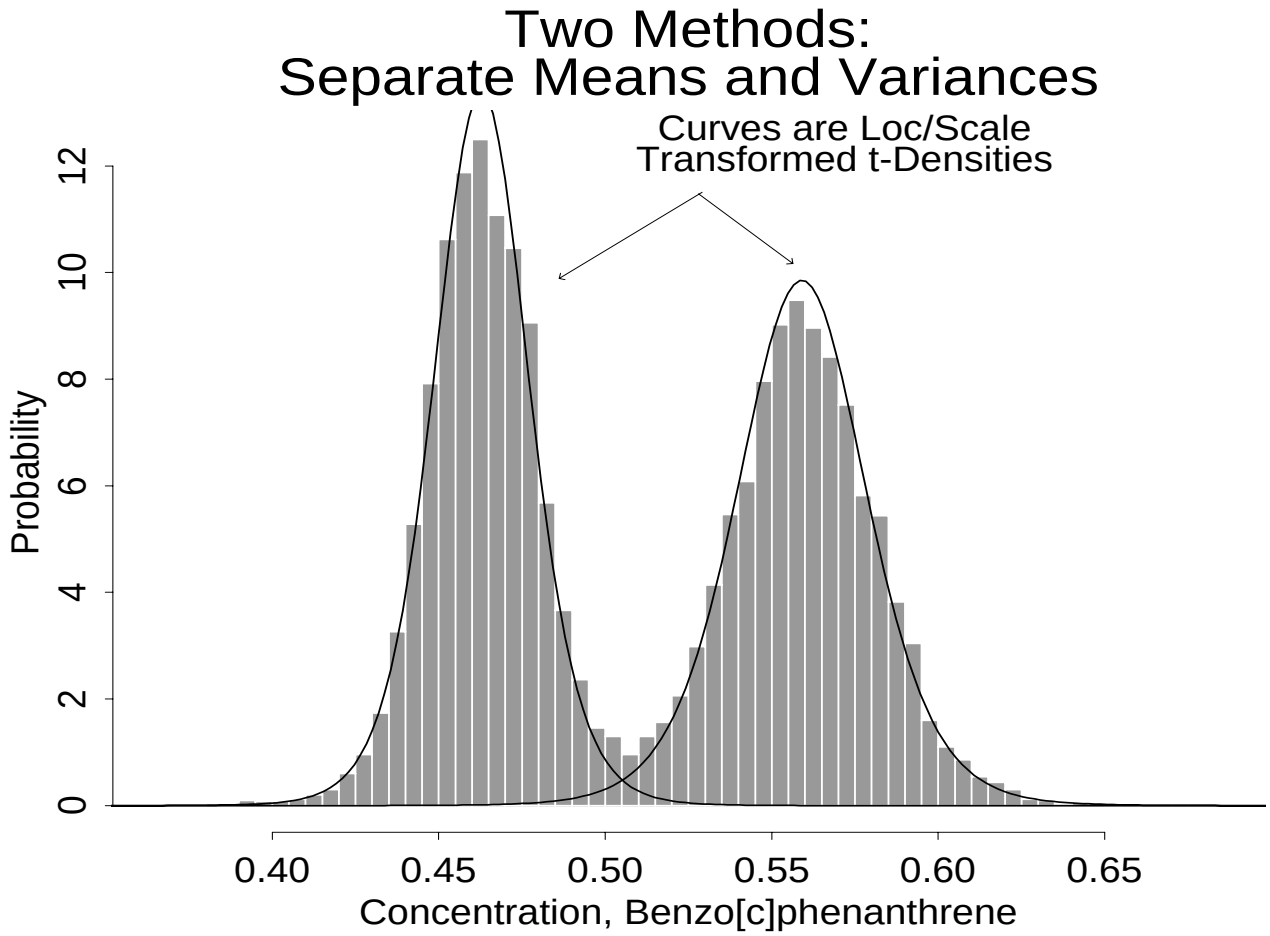
Given the model of the previous page, we can now calculate the posterior distributions of the method means θ_1 and θ_2 given the data $\left\{x_{ij}\right\}$:

$$p\left(\theta_i|\left\{x_{ij}\right\}\right) = \frac{\sqrt{n_i}}{s_i} T'_{n_i-1}\left(\frac{\theta_i - \bar{x}_i}{s_i/\sqrt{n_i}}\right)$$

Where $\bar{x}_i$, s_i and n_i are summary statistics for the ith method, and $T'(\cdot)_\nu$ denotes a t-density with ν degrees of freedom.

Here we can calculate the form of the posterior analytically. For more complicated situations we can simulate the posterior very accurately using Markov Chain Monte Carlo techniques and BUGS (Bayesian analysis Using Gibbs Sampling) software.

Model 1 Posterior for Benzo[c]phenanthrene



Here we see the posterior density functions of the method means θ_1 and θ_2 given the data $\left\{x_{ij}\right\}$.

Model 2:
Common Mean, Separate Variances
(Behrens-Fisher Problem)

$$\begin{aligned} p(x_{ij}|\theta, \sigma_i^2) &= \mathsf{N}(\theta, \sigma_i^2) \\ p(\theta) &\propto 1 \\ p(\sigma_i) &\propto 1/\sigma_i \end{aligned}$$

Here we model the observations from the different methods as coming from normal distributions with the same mean, but with different variances. The parameters of those normal distributions are again modeled with non-informative priors.

Model 2 Posterior: Poly-t Density

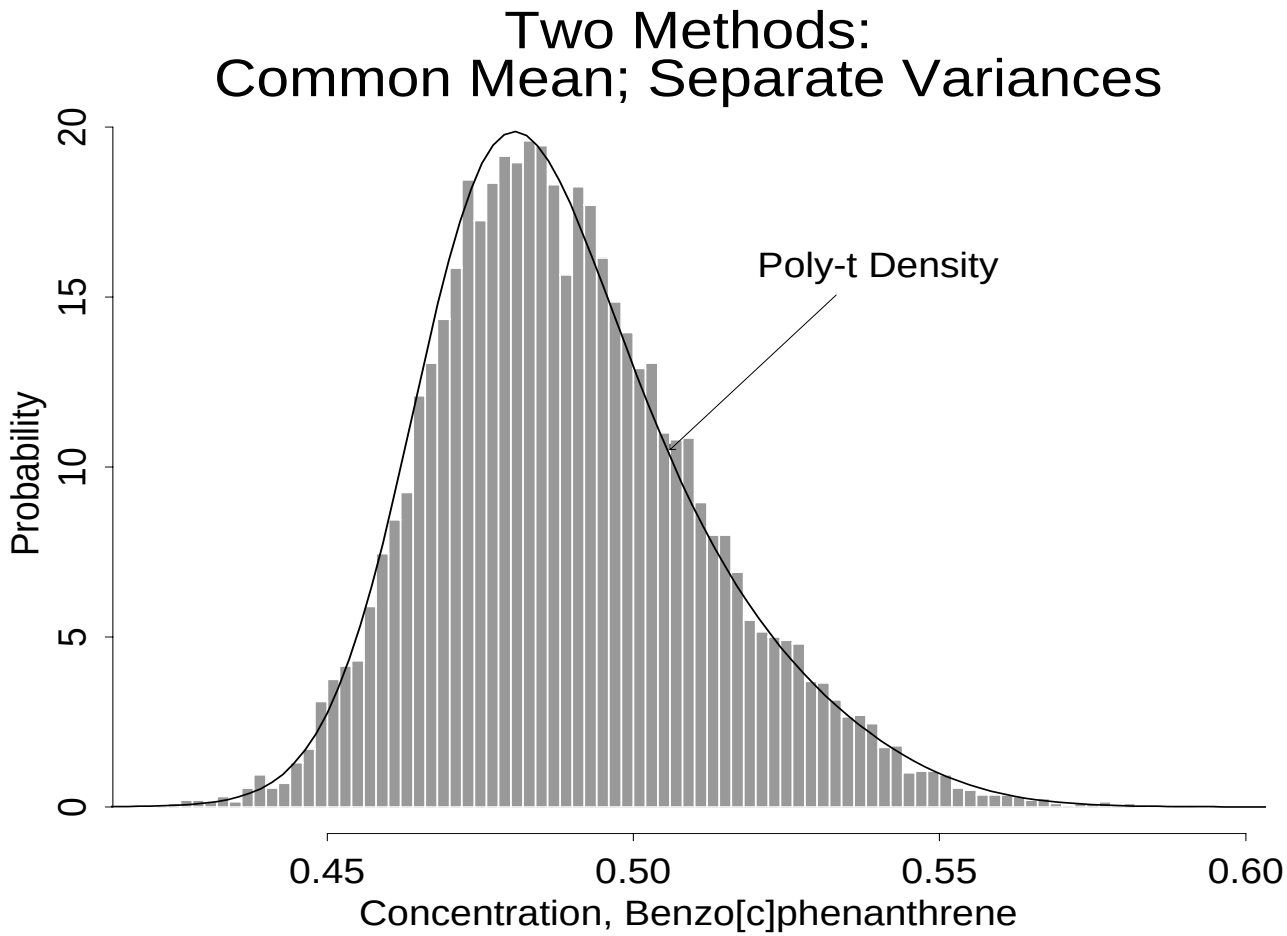
We can calculate analytically the posterior distribution of the common mean θ given the data $\left\{x_{ij}\right\}$:

$$p\left(\theta|\left\{x_{ij}\right\}\right) \propto \prod_{i=1}^{2} \frac{\sqrt{n_i}}{s_i} T'_{n_i-1}\left(\frac{\theta-\bar{x}_i}{s_i/\sqrt{n_i}}\right)$$

Where $\bar{x}_i$, s_i and n_i are summary statistics for the ith method, and $T'(\cdot)_\nu$ denotes a t-density with ν degrees of freedom.

For two methods, this posterior is related to the *Behrens-Fisher* distribution, but the result holds for an arbitrary number of methods [*Poly-t distribution*; See Box and Tiao, Bayesian Inference in Statistical Analysis (1973, Ch. 9)].

Model 2 Posterior for Benzo[c]phenanthrene



Model 3:
Two Levels for Mean, One Level for Variances

This is a hierarchical model where the method means are modeled to come from a normal prior distribution, and the parameters of that normal distribution have prior distributions of their own.

$$
\begin{aligned}
p(x_{ij}|\theta_i, \sigma_i^2) &= \mathsf{N}(\theta_i, \sigma_i^2) \\
p(\theta_i) &= \mathsf{N}(\mu, \sigma^2) \\
p(\sigma_i) &\propto 1/\sigma_i \\
p(\mu) &\propto 1 \\
p(\sigma) & \quad \text{“Arbitrary”}
\end{aligned}
$$

(Note that $p(\sigma)$ cannot blow up at $\sigma = 0$, since the likelihood will be nonzero there, and hence the posterior would be improper.)

Model 3 Posteriors: Generalization of Poly$-t$

Again we can analytically calculate the posteriors using t distributions:

$$p(\mu, \sigma | \{x_{ij}\}) \propto p(\sigma) \prod_{i=1}^{k} \frac{\sqrt{n_i}}{s_i} f_{n_i - 1} \left[\frac{\bar{x}_i - \mu}{s_i / \sqrt{n_i}}; \frac{2 n_i \sigma^2}{s_i^2} \right].$$

Where

$$f_\nu (u; \psi) \equiv \frac{1}{\Gamma_{\nu/2} \sqrt{\pi}} \int_0^\infty \frac{y^{(\nu+1)/2-1} e^{-y \left[1 + \frac{u^2}{\psi y + \nu} \right]}}{\sqrt{\psi y + \nu}} dy.$$

is the density of

$$U = T_\nu + Z \sqrt{\frac{\psi}{2}}$$

This result holds for k methods, not just for $k = 2$.

The Need for an Informative Prior on σ

- A frequentist is in serious trouble when the methods are regarded as 'random', since there is only one degree of freedom between methods.

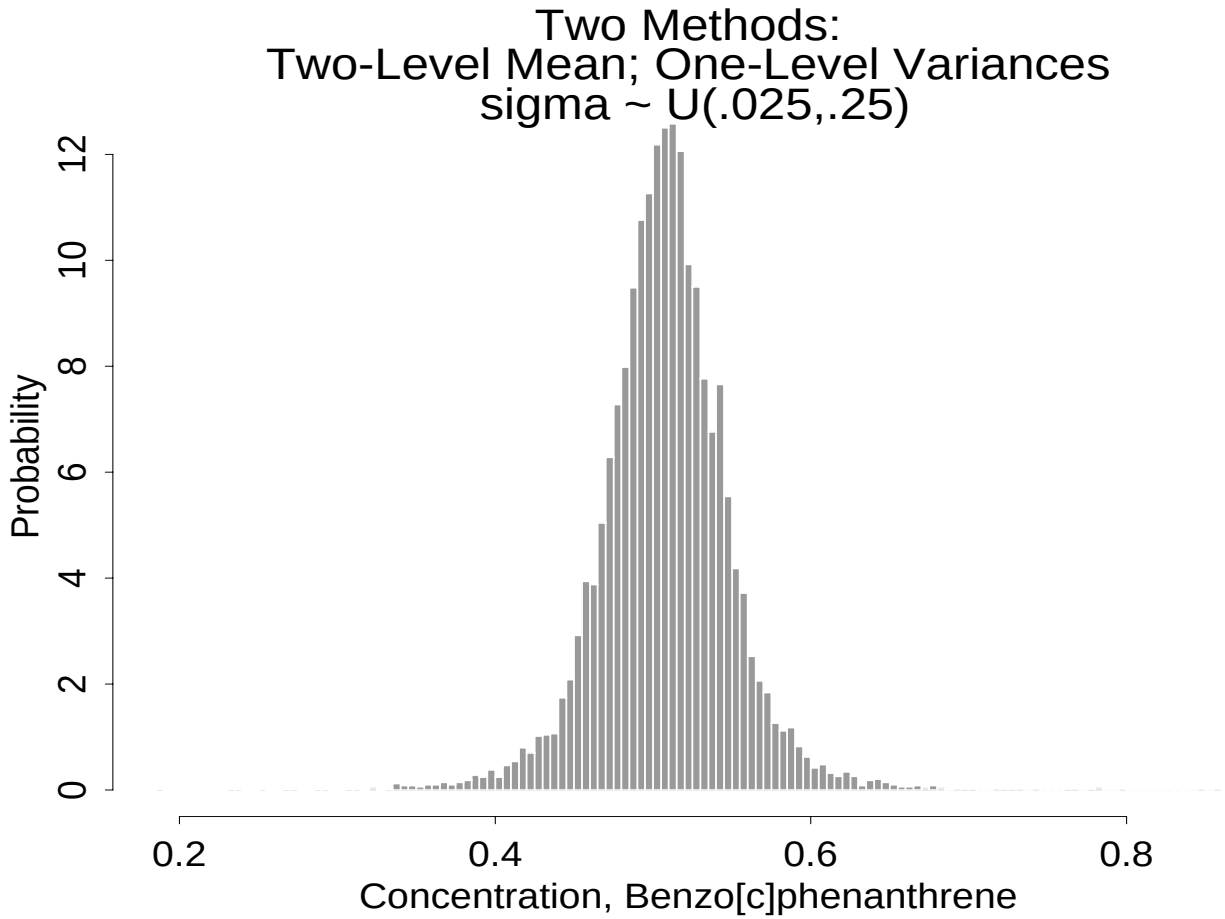
- A Bayesian trying to get 'objective' approximate frequentist results will have the same difficulty. (*Tanstaffl:* There's no such thing as a free lunch.)

- We need a proper prior on σ, and of course there are many 'reasonable' choices. For illustration, let

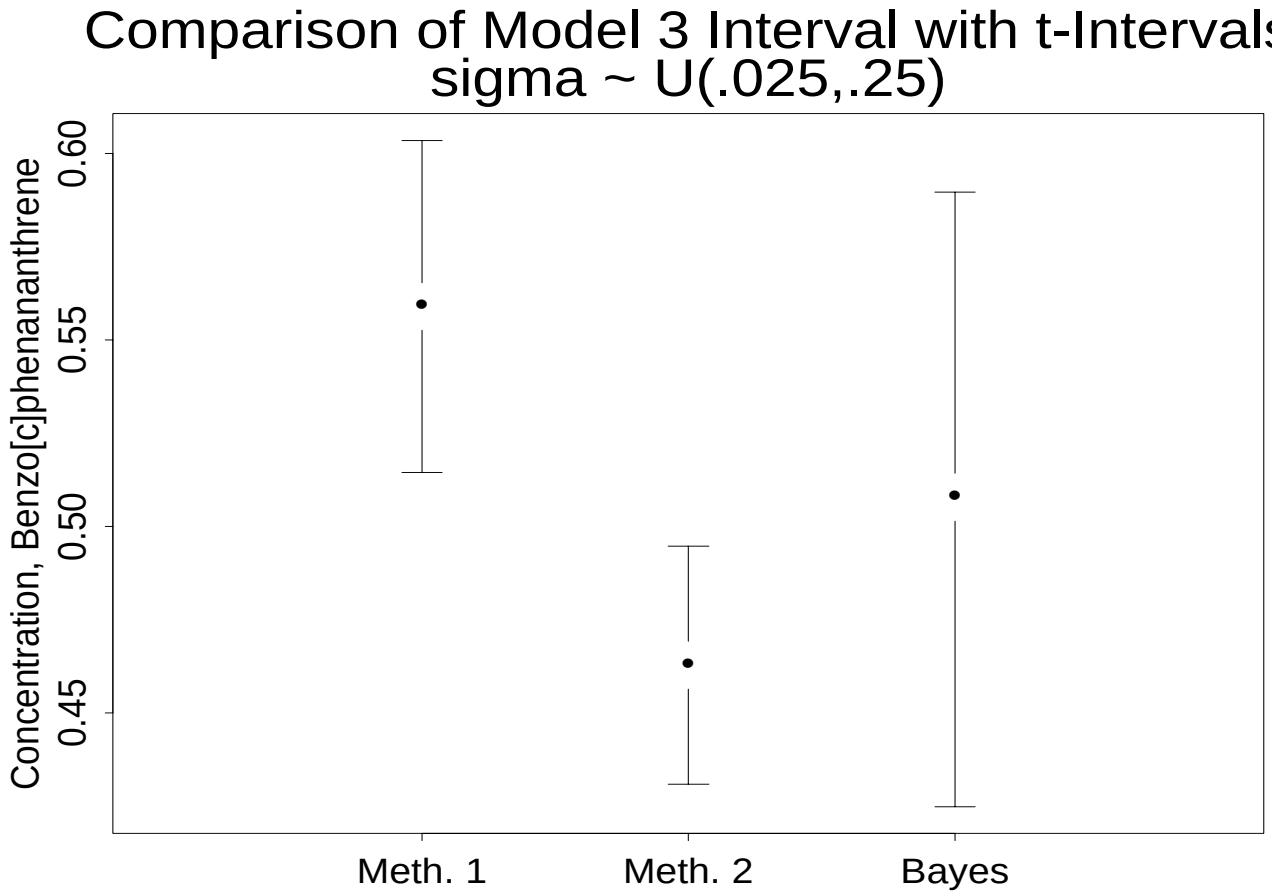
  $$p(\sigma) = U(.025, .25),$$

  the Uniform (Rectangular) distribution between .025 and .25.

Model 3 Posterior for Benzo[c]phenanthrene



Model 3 Bayesian Credible Interval



The Bayes interval in this case is large enough to encompass both method means.

BUGS Code for Model 3 Posterior Calculations

```
model twolvl;
const
  SAMPLES  = 22,
  METHODS  = 2;
var
  x[SAMPLES],  mth[SAMPLES],
  theta[METHODS], tau.within[METHODS],
  mu, tau.between;
data mth,x in "bcp.dat";
inits in "bcp.in";
{
for (i in 1:SAMPLES) {
      x[i] ~ dnorm(theta[mth[i]], tau.within[mth[i]]);
   }
for (i in 1:METHODS)  {
    tau.within[i] ~ dgamma(0.001, 0.001);
    theta[i]      ~ dnorm(mu, tau.between);
   }
tau.between ~ dunif(16,1600);
mu ~ dnorm(0, 1.E-10);
}
```